Using Web-scale N-grams to Improve Base NP Parsing Performance

Emily Pitler
University of Pennsylvania

Shane Bergsma
University of Alberta

Dekang Lin
Google, Inc.

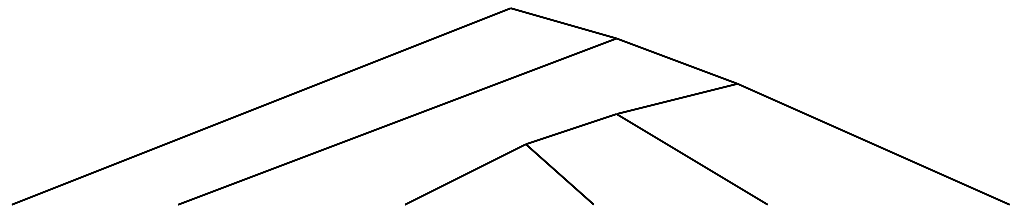
Kenneth Church
Johns Hopkins University

COLING 2010

Overview

- Goal: Recover the syntactic structure of arbitrary noun phrases (NPs):

its remaining mortgage loan origination operations

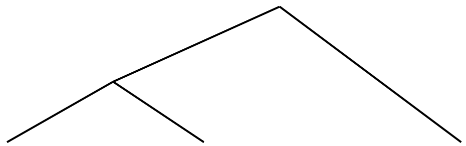


(its (remaining (((mortgage loan) origination) operations)))

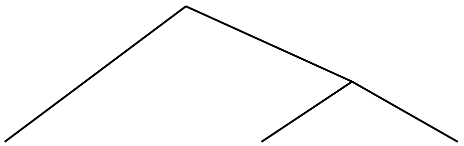
- Method: score all spans with classifier
- Features: statistics from web-scale N-grams
- Result: 95.4% exact match accuracy

Parsing NPs

- Classic 3-word cases: Left or Right?



(social science) teacher



retired (science teacher)

- Our objective: NPs of arbitrary length, including conjunctions:

  (French ((television and movie) producers))

Parsing NPs

- It's Hard: Number of binary trees increases with the Catalan numbers [Church & Patil, 1982]:

  2 (three words), **5, 14, 42** (six words), **132, 429, …**

- It's Worth Doing:

  70% of search-engine queries [Barr et al. 2008]

  (washed baby) carrots
  vs.
  washed (baby carrots)

Annotated Data

- Base NPs: NPs with no embedded NPs
- Penn Treebank originally had flat base NPs:

  (NP (JJ time-limited) (NN poison) (NNS pills))
- [Vadas and Curran, 2007] provide annotations for NPs in the WSJ articles from the Penn Treebank
  - 98.5% inter-annotator agreement accuracy

Examples

(((biotechnology research) (and (vaccine manufacturing))) concerns)
((The government) 's)
(executive (vice president))
(the (first time))
(its (four-year history))
(an (adverse (net-benefit decision)))
(the (same conclusions))
(industry (, (science (and technology))))
(Merieux (and Connaught))
(early (this week))
(the (government decision))
(Mehta (& Isaly))
((the government) 's)
(an (out-of-court settlement))
(a (settlement proposal))
(((research (and development)) spending) levels)
(Toronto-based ((Richardson Greenshields) Inc.))

Method: Span Scoring

- Supervised classifier (SVM) predicts probability of each span:

  (French television) *and movie producers*

  French ***(television and)*** *movie producers*

  (French television and) *movie producers*

  ...

  French television and ***(movie producers)***

- Can be feature in structured predictor [Taskar et al. 2004]

Method: Span Scoring

French television and movie producers

(French television)

(television and movie producers)

	French	television	and	movie	producers
French		.15	.07	.37	1
television			.14	.51	.76
and				.78	.33
movie					.25
producers					

(French ((television (and movie)) producers)))

Prior work leveraging raw text

"retired science teacher"

- Adjacency model [Marcus, 1980; Liberman and Sproat, 1992; Pustejovsky et al., 1993; Resnik, 1993]
  - "retired science" vs. "science teacher" (e.g. PMI)
- Dependency model [Lauer, 1995]
  - "retired science" vs. "retired teacher" (e.g. PMI)
- We generalize these models for features for longer NPs (counts from web-scale N-grams)

PMIs for "movie producers"

e.g. "French television and **movie producers**"

- PMI used as association measure

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

- Features for all pairs:

a) High PMI(movie,producers) → positive evidence

b) ↑PMI(and,movie) → negative evidence

c) ↑PMI(television,and) → slightly-positive evidence

PMIs for "movie producers"

French	television
French	and
French	movie
French	producers
television	and
television	movie

television	producers
and	movie
and	producers
movie	producers
television and	and movie
television and	and producers

Special Conjunction Features

- How to tell "television and movie" go together?
  - PMI(television,and), PMI(and,movie), PMI(television,movie) are insufficient
- Special PMI_{and} features:

$\text{PMI}_{\text{and}}(\text{television,movie}) =$

$$\log\left(\frac{p(\text{"television and movie"})}{p(\text{"television and"})\,p(\text{"and movie"})}\right)$$

Method: Span Scoring

French television and movi

(movie producers) is usually a good bet, but has low probability in this context

	French	television	and	mov	producers
French		.15	.07	.37	1
television			.14	.51	.76
and				.78	.33
movie					.25
producers					

(French ((television (and movie)) producers)))

Web-Scale N-gram Data

- Details in: [Lin et al., LREC 2010]
  - Same source as Google N-grams Version 1
  - More pre-processing: duplicate sentence removal, length+alphabetical constraints
  - Fast lookup tools based on suffix arrays

```
time cat phrases.txt | multi_lookup /export/ws09/dlin/church/Google/V2
```

```
real    21m50.273s
user    0m1.131s
sys     0m46.021s
```

168K phrases

Non N-gram Features

- Lexical features
  - Word at each position
- “Shape” features
  - Captures upper and lower case, punctuation
- Position feature
  - Prior probability of bracketing at that position

Experiments

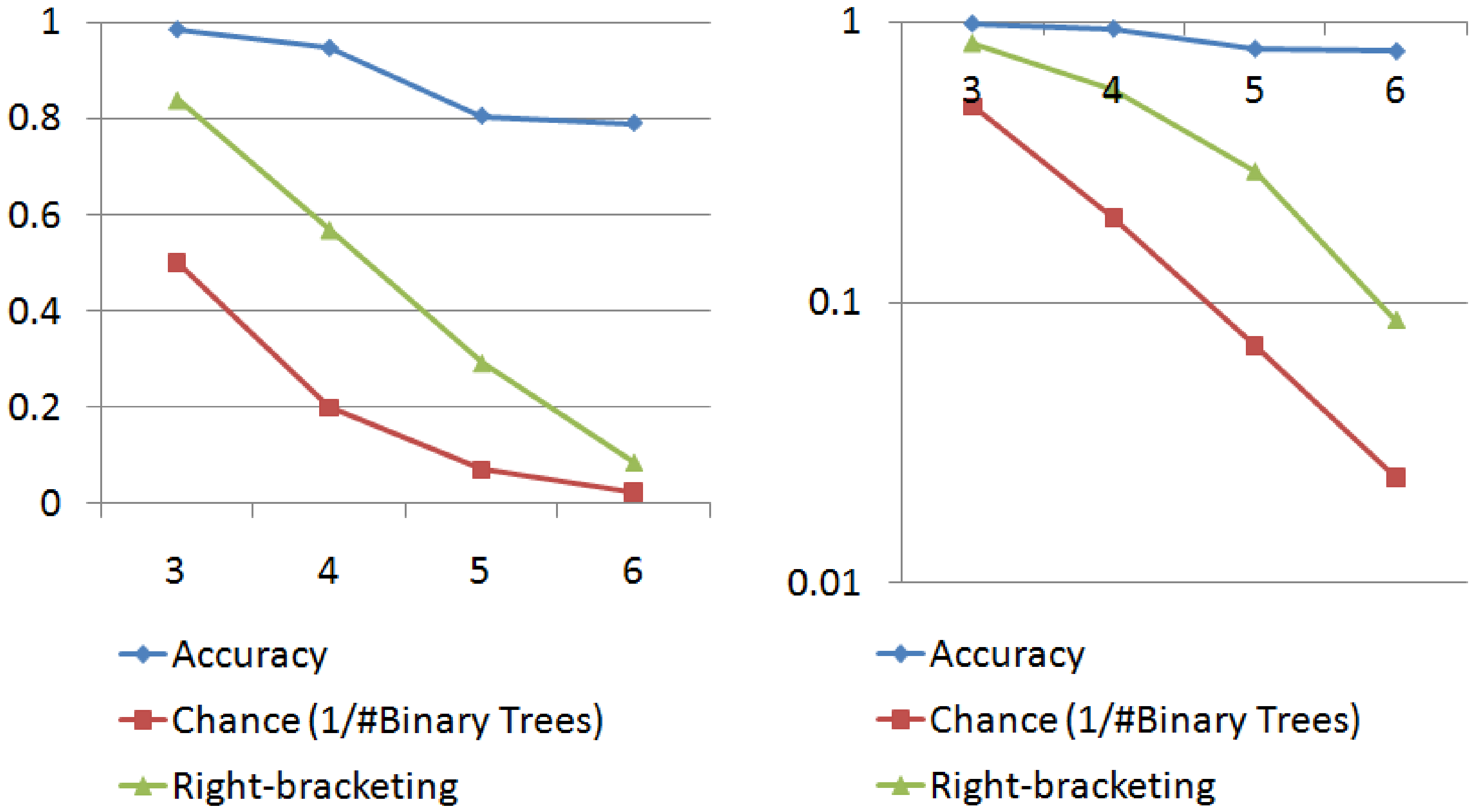
- Standard splits of WSJ data, using annotations from [Vadas & Curran, 2007]
- All >2-word Base NPs
- Report **accuracy (%)**

Results

Method	Accuracy
Right-bracketing baseline	72.6
Lex, Shape, Position Features	94.0
+N-gram PMI features	**95.4**
[Vadas & Curran, 2007b]	93.0*

*not really comparable

Accuracy by NP length



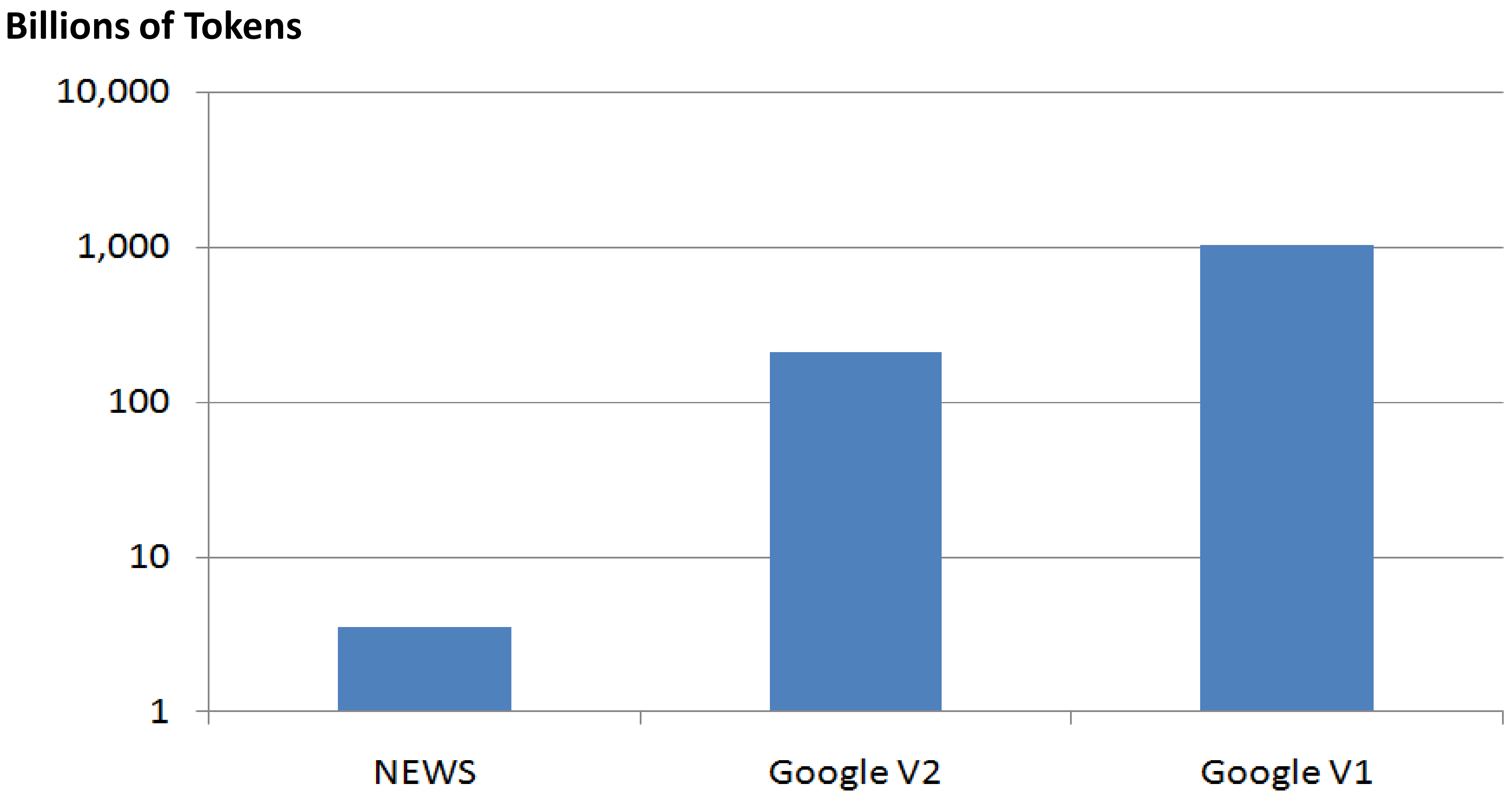
N-gram data helps most on conjunctions

Method	Conjunctions	Everything else
Lex, Shape, Position	84.0	94.5
+Ngrams	**89.7 (+6%)**	**95.7 (+1%)**

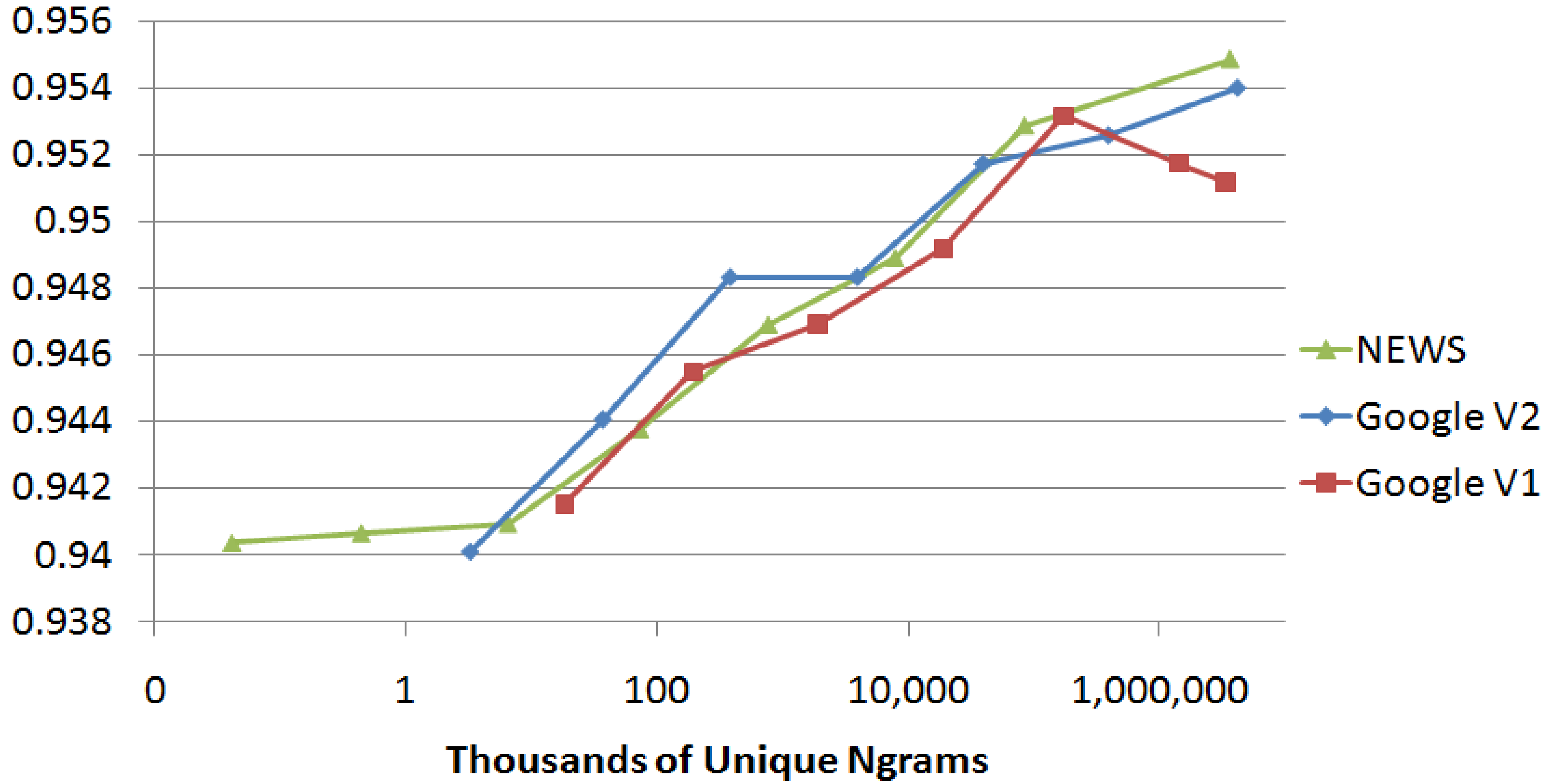
Less Data is Worse Data

N-gram Cut-off	# Unique N-grams	Accuracy
10	4,145,972,000	95.40%
100	391,344,991	95.30%
1,000	39,368,488	95.20%
10,000	3,924,478	94.80%
100,000	386,639	94.80%
1,000,000	37,567	94.40%
10,000,000	3,317	94.00%

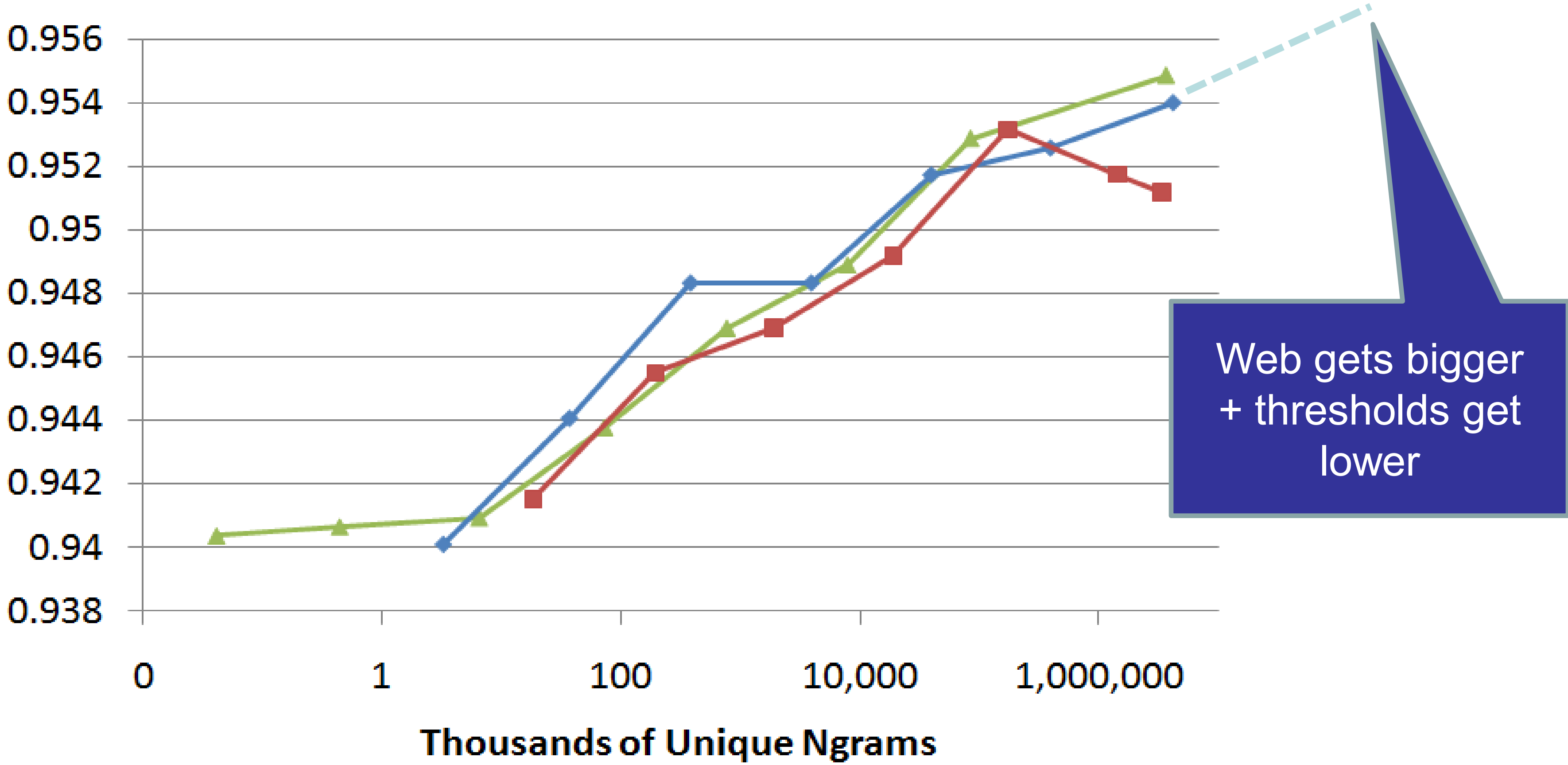
What about other data?



Number of Unique N-grams!



Number of Unique N-grams!



Conclusion

- New standard in Base NP parsing performance: 95.4%
- N-gram data particularly helps on conjunctions
- Log-linear gain with amount of UNLABELED data (number of unique N-grams)

Thanks

- Center for Language & Speech Processing at Johns Hopkins University
- Our colleagues on *Team N-gram*